Supplementary material: Finding genes in *Schistosoma japonicum*: Annotating novel genomes with help of extrinsic evidence

January 12, 2009

Table S1: **Evaluation of iterative training on *C. elegans*.** Using supported gene fragments helps iterative training to achieve the performance close to that of the training on curated training set. Filtering step is important in achieving high performance when relying on cross-species evidence.

		Gene		**Exon**		**InternalE**		**Intron**		**Nucleotide**	
Training	Iter.	sn	sp	sn	sp	sn	sp	sn	sp	sn	sp
All evidence (both training and testing)											
human		25%	26%	61%	72%	67%	79%	67%	79%	88%	91%
c.elegans		45%	46%	82%	80%	90%	82%	88%	83%	96%	92%
iterative/supported	1	44%	44%	81%	79%	89%	82%	87%	82%	96%	92%
iterative/supported	2	44%	45%	82%	78%	89%	81%	88%	82%	96%	92%
iterative/supported	3	43%	44%	82%	78%	89%	81%	88%	81%	96%	92%
iterative/unfiltered	1	43%	45%	81%	80%	89%	82%	88%	83%	96%	93%
iterative/unfiltered	2	40%	42%	81%	79%	90%	81%	88%	82%	96%	93%
iterative/unfiltered	3	40%	42%	81%	79%	90%	81%	88%	82%	96%	93%
Cross-species evidence (both training and testing)											
human		5%	8%	41%	54%	48%	60%	43%	55%	81%	86%
c.elegans		27%	33%	76%	75%	87%	77%	84%	78%	96%	92%
iterative/supported	1	23%	27%	72%	72%	82%	75%	80%	77%	94%	92%
iterative/supported	2	24%	27%	73%	72%	84%	74%	81%	75%	94%	91%
iterative/supported	3	24%	28%	73%	72%	84%	75%	81%	76%	94%	91%
iterative/unfiltered	1	17%	24%	71%	73%	83%	76%	79%	78%	94%	93%
iterative/unfiltered	2	17%	23%	72%	72%	84%	75%	81%	77%	94%	93%
iterative/unfiltered	3	17%	23%	72%	72%	84%	75%	81%	77%	94%	93%
Ab initio (both training and testing)											
human		2%	4%	36%	50%	43%	56%	37%	51%	77%	86%
c.elegans		23%	30%	76%	74%	87%	76%	84%	78%	95%	92%
iterative/unfiltered	1	12%	19%	68%	72%	81%	75%	77%	77%	93%	94%
iterative/unfiltered	2	15%	20%	71%	72%	84%	75%	81%	77%	94%	93%
iterative/unfiltered	3	15%	20%	71%	72%	84%	75%	81%	77%	94%	93%

Table S2: **Effectiveness of iterative training in *ab initio* gene finding.** The table compares performance of *ab initio* gene finding in *C. elegans* genome using different variants of iterative training. Using evidence and filtering step in iterative training is instrumental in achieving performance close to that of training on curated training set.

		Gene		**Exon**		**InternalE**		**Intron**		**Nucleotide**	
Training	Iter.	sn	sp	sn	sp	sn	sp	sn	sp	sn	sp
All evidence (training only)											
iterative/supported	2	22%	28%	75%	72%	87%	74%	83%	76%	95%	91%
iterative/unfiltered	2	18%	25%	74%	73%	87%	75%	83%	77%	95%	93%
Cross-species evidence (training only)											
iterative/supported	2	22%	26%	72%	71%	83%	74%	80%	75%	94%	91%
iterative/unfiltered	2	14%	20%	71%	72%	84%	74%	81%	77%	94%	93%
Ab initio											
iterative/unfiltered	2	15%	20%	71%	72%	84%	75%	81%	77%	94%	93%
Baseline											
human		2%	4%	36%	50%	43%	56%	37%	51%	77%	86%
c.elegans		23%	30%	76%	74%	87%	76%	84%	78%	95%	92%

Table S3: **Effectiveness of iterative training for gene finding with evidence.** The table compares performance of gene finding with evidence on *C. elegans* testing set using different variants of iterative training. Although the use of evidence in testing increases the overall accuracy, the new iterative training on supported predictions leads to better gene-level accuracy than training on unfiltered data.

		Gene		**Exon**		**Int. exon**		**Intron**		**Nucleotide**	
Training	Iter.	sn	sp	sn	sp	sn	sp	sn	sp	sn	sp
All evidence (training and testing)											
iterative/supported	3	43%	44%	82%	78%	89%	81%	88%	81%	96%	92%
iterative/unfiltered	3	39%	41%	81%	78%	90%	81%	88%	82%	96%	93%
Cross-species evidence training, all evidence testing											
iterative/supported	3	41%	41%	81%	78%	88%	81%	86%	81%	95%	92%
iterative/unfiltered	3	36%	39%	80%	77%	89%	80%	87%	81%	96%	93%
Ab initio training, all evidence testing											
iterative/unfiltered	3	36%	38%	80%	77%	89%	80%	87%	81%	96%	93%
Baseline											
human		25%	26%	61%	72%	67%	79%	67%	79%	88%	91%
C.elegans		45%	46%	82%	80%	90%	82%	88%	83%	96%	92%